SVD and Low-rank Approximation

Lecture 23

April 18, 2019

Singular Value Decomposition (SVD)

Let A be a $m \times n$ real-valued matrix

- a_i denotes vector corresponding to row i
- m rows. think of each row as a data point in $\mathbb{R}^n$
- Data applications: $m \gg n$
- Other notation: A is a $n \times d$ matrix.

Singular Value Decomposition (SVD)

Let A be a $m \times n$ real-valued matrix

- a_i denotes vector corresponding to row i
- m rows. think of each row as a data point in $\mathbb{R}^n$
- Data applications: $m \gg n$
- Other notation: A is a $n \times d$ matrix.

SVD theorem: A can be written as UDV^T where

- V is a $n \times n$ orthonormal matrix
- D is a $m \times n$ diagonal matrix with $\leq \min\{m, n\}$ non-zeroes called the singular values of A
- U is a $m \times m$ orthonormal matrix

SVD

Let $d = \min\{m, n\}$.

- $u_1, u_2, \ldots, u_m$ columns of U, left singular vectors of A
- $v_1, v_2, \ldots, v_n$ columns of V (rows of V^T) right singular vectors of A
- $\sigma_1 \geq \sigma_2 \geq \ldots, \geq \sigma_d$ are singular values where $d = \min\{m, n\}$. And $\sigma_i = D_{i,i}$

$$A = \sum_{i=1}^{d} \sigma_i u_i v_i^T$$

SVD

Let $d = \min\{m, n\}$.

- $u_1, u_2, \ldots, u_m$ columns of U, left singular vectors of A
- $v_1, v_2, \ldots, v_n$ columns of V (rows of V^T) right singular vectors of A
- $\sigma_1 \geq \sigma_2 \geq \ldots, \geq \sigma_d$ are singular values where $d = \min\{m, n\}$. And $\sigma_i = D_{i,i}$

$$A = \sum_{i=1}^{d} \sigma_i u_i v_i^T$$

We can in fact restrict attention to r the rank of A.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

SVD

Interpreting A as a linear operator $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$

- Columns of V is an orthonormal basis and hence $V^T x$ for $x \in \mathbb{R}^n$ expresses x in the V basis. Note that $V^T x$ is a rigid transformation (does not change length of x).
- Let $y = V^T z$. D is a diagonal matrix which only stretches y along the coordinate axes. Also adjusts dimension to go from n to m with right number of zeroes.
- Let $z = Dy$. Then Uz is a rigid transformation that expresses z in the basis corresponding to rows of U.

Thus any linear operator can be broken up into a sequence of three simpler/basic type of transformations

Low rank approximation property of SVD

Question: Given $A \in \mathbb{R}^{m \times n}$ and integer k find a matrix B of rank at most k such that $\|A - B\|$ is minimized

Low rank approximation property of SVD

Question: Given $A \in \mathbb{R}^{m \times n}$ and integer k find a matrix B of rank at most k such that $\|A - B\|$ is minimized

Fact: For Frobenius norm optimum for all k is captured by SVD.

That is, $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ is the best rank k approximation to A

$$\|A - A_k\|_F = \min_{B:\text{rank}(B) \leq k} \|A - B\|_F$$

Low rank approximation property of SVD

Question: Given $A \in \mathbb{R}^{m \times n}$ and integer k find a matrix B of rank at most k such that $\|A - B\|$ is minimized

Fact: For Frobenius norm optimum for all k is captured by SVD.

That is, $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ is the best rank k approximation to A

$$\|A - A_k\|_F = \min_{B:\text{rank}(B) \leq k} \|A - B\|_F$$

Why this magic? Frobenius norm and basic properties of vector projections

Geometric meaning

Consider $k = 1$. What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Wlog v is a unit vector

Geometric meaning

Consider $k = 1$. What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Wlog v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

Geometric meaning

Consider $k = 1$. What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Wlog v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

If we know v then best u to minimize above is determined. Why?

Geometric meaning

Consider $k = 1$. What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Wlog v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

If we know v then best u to minimize above is determined. Why?
For fixed v, $u(i) = \langle a_i, v \rangle$

Geometric meaning

Consider $k = 1$. What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Wlog v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

If we know v then best u to minimize above is determined. Why?
For fixed v, $u(i) = \langle a_i, v\rangle$
$\|a_i - \langle a_i, v\rangle v\|_2$ is distance of a_i from line described by v.

Geometric meaning

What is the best rank $\mathbf{1}$ matrix B that minimizes $\|A - B\|_F$

It is to find unit vector/direction v to minimize

$$\sum_{i=1}^{m} ||a_i - \langle a_i, v \rangle v||^2$$

which is same as finding unit vector v to maximize

$$\sum_{i=1}^{m} \langle a_i, v \rangle^2$$

Geometric meaning

What is the best rank $\mathbf{1}$ matrix B that minimizes $\|A - B\|_F$

It is to find unit vector/direction v to minimize

$$\sum_{i=1}^{m} ||a_i - \langle a_i, v \rangle v||^2$$

which is same as finding unit vector v to maximize

$$\sum_{i=1}^{m} \langle a_i, v \rangle^2$$

How to find best v? Not obvious: we will come to it a bit later

Best rank two approximation

Consider $k = 2$. What is the best rank 2 matrix B that minimizes $\|A - B\|_F$

Since B has rank 2 we can assume without loss of generality that $B = u_1 v_1^T + u_2 v_2^T$ where v_1, v_2 are orthogonal unit vectors (span a space of dimension 2)

Best rank two approximation

Consider $k = 2$. What is the best rank 2 matrix B that minimizes $\|A - B\|_F$

Since B has rank 2 we can assume without loss of generality that $B = u_1 v_1^T + u_2 v_2^T$ where v_1, v_2 are orthogonal unit vectors (span a space of dimension 2)

Minimizing $\|A - B\|_F^2$ is same as finding orthogonal vectors v_1, v_2 to maximize

$$\sum_{i=1}^{m} (\langle a_i, v_1 \rangle^2 + \langle a_i, v_2 \rangle^2)$$

in other words the best fit 2-dimensional space

Greedy algorithm

- Find v_1 as the best rank 1 approximation. That is $v_1 = \arg\max_{v, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$
- For v_2 solve $\arg\max_{v \perp v_1, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$.

  Alternatively: let $a_i' = a_i - \langle a_i, v_1 \rangle v_1$. Let $v_2 = \arg\max_{v, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a_i', v \rangle^2$

Greedy algorithm

- Find v_1 as the best rank 1 approximation. That is $v_1 = \arg\max_{v, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$
- For v_2 solve $\arg\max_{v \perp v_1, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$.

  Alternatively: let $a'_i = a_i - \langle a_i, v_1 \rangle v_1$. Let $v_2 = \arg\max_{v, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a'_i, v \rangle^2$

Greedy algorithm works!

Greedy algorithm correctness

Proof that Greedy works for $k = 1$.

Suppose w_1, w_2 are orthogonal unit vectors that form the best fit 2-d space. Let H be the space spanned by w_1, w_2.

Suffices to prove that

$$\sum_{i=1}^{m}(\langle a_i, v_1\rangle^2 + \langle a_i, v_2\rangle^2) \geq \sum_{i=1}^{m}(\langle a_i, w_1\rangle^2 + \langle a_i, w_2\rangle^2)$$

Greedy algorithm correctness

Proof that Greedy works for $k = 1$.

Suppose w_1, w_2 are orthogonal unit vectors that form the best fit 2-d space. Let H be the space spanned by w_1, w_2.

Suffices to prove that

$$\sum_{i=1}^{m}(\langle a_i, v_1\rangle^2 + \langle a_i, v_2\rangle^2) \geq \sum_{i=1}^{m}(\langle a_i, w_1\rangle^2 + \langle a_i, w_2\rangle^2)$$

If $v_1 \subset H$ then done because we can assume wlog that $w_1 = v_1$ and v_2 is at least as good as w_2.

Greedy algorithm correctness

Suppose $v_1 \not\in H$. Let v_1' be projection of v_1 onto H and $v_1'' = v_1 - v_1'$ be the component of v_1 orthogonal to H.

Greedy algorithm correctness

Suppose $v_1 \not\in H$. Let v_1' be projection of v_1 onto H and $v_1'' = v_1 - v_1'$ be the component of v_1 orthogonal to H. Note that $\|v_1'\|^2 + \|v_1''\|_2^2 = \|v_1\|_2^2 = 1$.

Wlog we can assume by rotation that $w_1 = \frac{1}{\|v_1'\|_2} v_1'$ and w_2 is orthogonal to v_1'. Hence w_2 is orthogonal to v_1.

Greedy algorithm correctness

Suppose $\boldsymbol{v_1} \not\in \boldsymbol{H}$. Let $\boldsymbol{v_1'}$ be projection of $\boldsymbol{v_1}$ onto $\boldsymbol{H}$ and $\boldsymbol{v_1''} = \boldsymbol{v_1} - \boldsymbol{v_1'}$ be the component of $\boldsymbol{v_1}$ orthogonal to $\boldsymbol{H}$. Note that $\|\boldsymbol{v_1'}\|^2 + \|\boldsymbol{v_1''}\|_2^2 = \|\boldsymbol{v_1}\|_2^2 = \boldsymbol{1}$.

Wlog we can assume by rotation that $\boldsymbol{w_1} = \frac{1}{\|\boldsymbol{v_1'}\|_2}\boldsymbol{v_1'}$ and $\boldsymbol{w_2}$ is orthogonal to $\boldsymbol{v_1'}$. Hence $\boldsymbol{w_2}$ is orthogonal to $\boldsymbol{v_1}$.

Therefore $\boldsymbol{v_2}$ is at least as good as $\boldsymbol{w_2}$, and $\boldsymbol{v_1}$ is at least as good as $\boldsymbol{w_1}$ which implies the desired claim.

Greedy algorithm for general k

- Find v_1 as the best rank 1 approximation. That is $v_1 = \arg\max_{v, \|v\|_2=1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$
- For v_k solve $\arg\max_{v \perp v_1, v_2, \ldots, v_{k-1}, \|v\|_2=1} \sum_{i=1}^{k} \langle a_i, v \rangle^2$ which is same as solving $k = 1$ with vectors $a'_1, a'_2, \ldots, a'_m$ that are residuals. That is $a'_i = a_i - \sum_{j=1}^{k-1} \langle a_i, v_j \rangle v_j$

Proof of correctness is via induction and is a straight forward generalization of the proof for $k = 2$

Summarizing

$$\sigma_j^2 = \sum_{i=1}^{m} \langle a_i, v_j \rangle^2$$

By greedy contruction $\sigma_1 \geq \sigma_2 \ldots,$

Let r be the (row) rank of A. $v_1, v_2, \ldots, v_r$ span the row space of A and $\sigma_j = 0$ for $j > r$

u_1 determined by v_1 and u_2 determined by v_1, v_2 and so on. Can show that they are orthogonal.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Power method

Thus SVD relies on being able to solve $k = 1$ case

Given m vectors $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ solve

$$\max_{v \in \mathbb{R}^n, \|v\|_2 = 1} \langle a_i, v \rangle^2$$

How do we solve the above problem?

Let $B = A^T A$ Then

$$\begin{aligned} B &= (\sum_{i=1}^{m} \sigma_i v_i u_i^T)(\sum_{i=1}^{r} \sigma_i u_i v_i^T) \\ &= \sum_{i=1}^{r} \sigma_i^2 v_i v_i^T \end{aligned}$$

Power method continued

Let $B = A^T A$ Then

$$\begin{aligned} B^2 &= (\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T)(\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T) \\ &= \sum_{i=1}^{r} \sigma_i^4 v_i v_i^T. \end{aligned}$$

More generally

$$B^k = \sum_{i=1}^{r} \sigma_i^k v_i v_i^T$$

Power method continued

Let $B = A^T A$ Then

$$\begin{aligned} B^2 &= (\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T)(\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T) \\ &= \sum_{i=1}^{r} \sigma_i^4 v_i v_i^T. \end{aligned}$$

More generally

$$B^k = \sum_{i=1}^{r} \sigma_i^k v_i v_i^T$$

If $\sigma_1 > \sigma_2$ then B^k converges to $\sigma_1^k v_1 v_1^T$ and we can identify v_1 from B^k. But expensive to compute B^k

Power method continued

Pick a random (unit) vector $x \in \mathbb{R}^n$. Then $x = \sum_{i=1}^{n} \lambda_i v_i$ since $v_1, v_2, \ldots, v_n$ is a basis for $\mathbb{R}^n$.

$$B^k x = (\sum_{i=1}^{r} \sigma_i^k v_i v_i^T)(\sum_{i=1}^{d} \lambda_i v_i) \to \sigma_1^{2k} \lambda_1 v_1$$

Can obtain v_1 by normalizing $B^k x$ to a unit vector.
Computing $B^k x$ is easier via a series of matrix vector multiplications

Power method continued

Pick a random (unit) vector $x \in \mathbb{R}^n$. Then $x = \sum_{i=1}^{n} \lambda_i v_i$ since $v_1, v_2, \ldots, v_n$ is a basis for $\mathbb{R}^n$.

$$B^k x = (\sum_{i=1}^{r} \sigma_i^k v_i {v_i}^T)(\sum_{i=1}^{d} \lambda_i v_i) \to \sigma_1^{2k} \lambda_1 v_1$$

Can obtain v_1 by normalizing $B^k x$ to a unit vector.
Computing $B^k x$ is easier via a series of matrix vector multiplications

Why random x?

What if $\sigma_1 \simeq \sigma_2$? Power method still works. See references.

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Interesting when $m > n$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Interesting when $m > n$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.

Geometrically Ax is a linear combination of columns of A. Hence we are asking what is the vector z in the column space of A that is closest to vector b in ℓ_2 norm.

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Interesting when $m > n$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.

Geometrically Ax is a linear combination of columns of A. Hence we are asking what is the vector z in the column space of A that is closest to vector b in ℓ_2 norm.

Closest vector to b is the projection of b into the column space of A so it is "obvious" geometrically. How do we find it?

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Interesting when $m > n$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.

Geometrically Ax is a linear combination of columns of A. Hence we are asking what is the vector z in the column space of A that is closest to vector b in ℓ_2 norm.

Closest vector to b is the projection of b into the column space of A so it is "obvious" geometrically. How do we find it? Find an orthonormal basis $z_1, z_2, \ldots, z_r$ for the columns of A. Compute projection b' as $b' = \sum_{j=1}^{r} \langle b, z_j \rangle z_j$ and output answer as $\|b - b'\|_2$.

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Closest vector to b is the projection of b into the column space of A so it is "obvious" geometrically. Find an orthonormal basis $z_1, z_2, \ldots, z_r$ for the columns of A. Compute projection b' as $b' = \sum_{j=1}^{r} \langle b, z_j \rangle z_j$ and output answer as $\|b - b'\|_2$.

Finding the basis is the expensive part. Recall SVD gives $v_1, v_2, \ldots, v_r$ which form a basis for the *row* space of A but then $u_1^T, u_2^T, \ldots, u_m^T$ form a basis for the *column* space of A. Hence SVD gives us all the information to find b'. In fact we have

$$\min_x \|Ax - b\|_2^2 = \sum_{i=r+1}^{m} \langle u_i^T, b \rangle^2$$